The Thermal Wall:

Where it came from and how to live with it?

Ronny Ronen
Senior Principal Engineer
Director of Architecture Research
Microprocessor Technology Labs
Haifa, Israel

Intel Corporation

The 10th Intel
EMEA Academic forum
Gdansk, Poland
May 19, 2005

Contributors: Lev Finkelstein, Eli Savransky, Efi Rotem

Point Mugu (USA) Airshow 2004

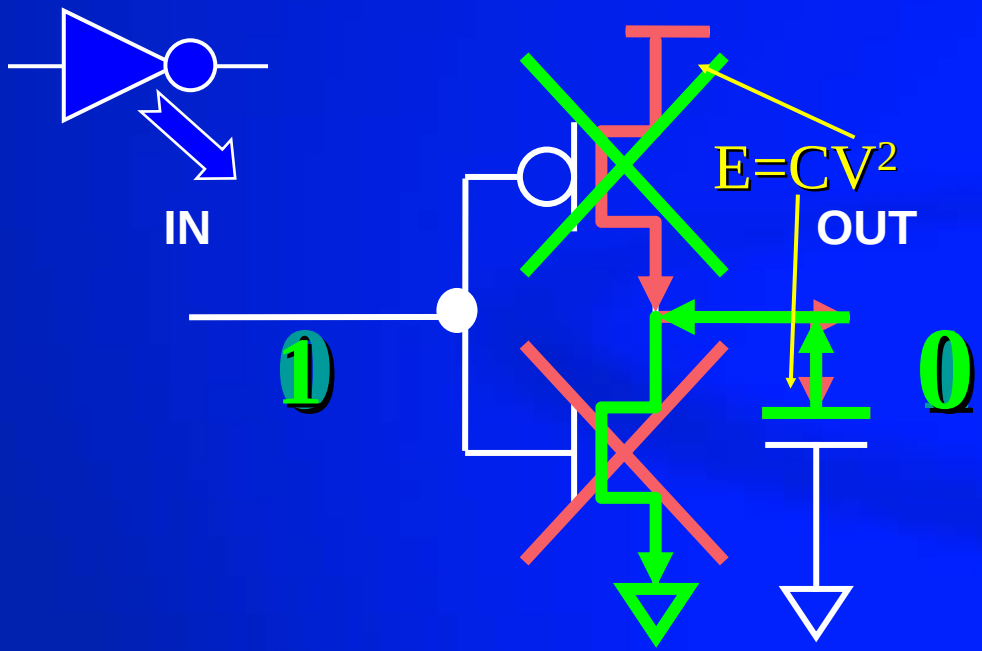
http://www.richard-seaman.com/Aircraft/AirShows/PointMugu2004/Highlights/

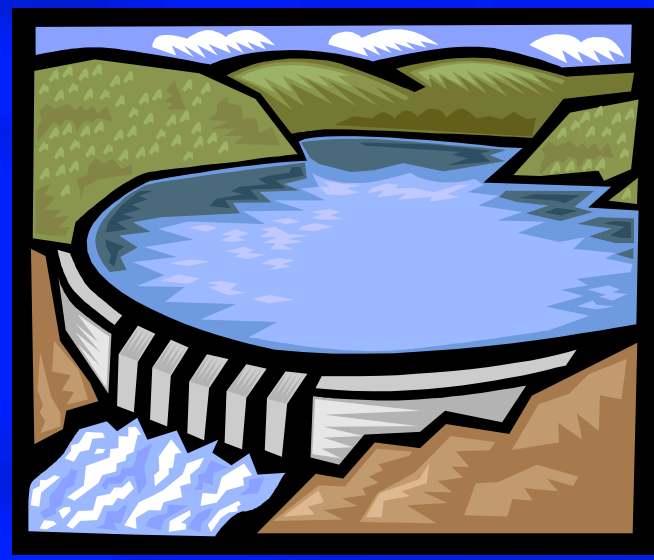
Agenda

- Background
- Process Scaling
- The Thermal Challenge
- Variations
- Dynamic Thermal Management
- Conclusions

Power and the digital world (1)…

- Power is consumed:
  - **When capacitance is charged and discharged.**
  - **A charged cap is a logical ‘1’, a discharged cap is ‘0’.**

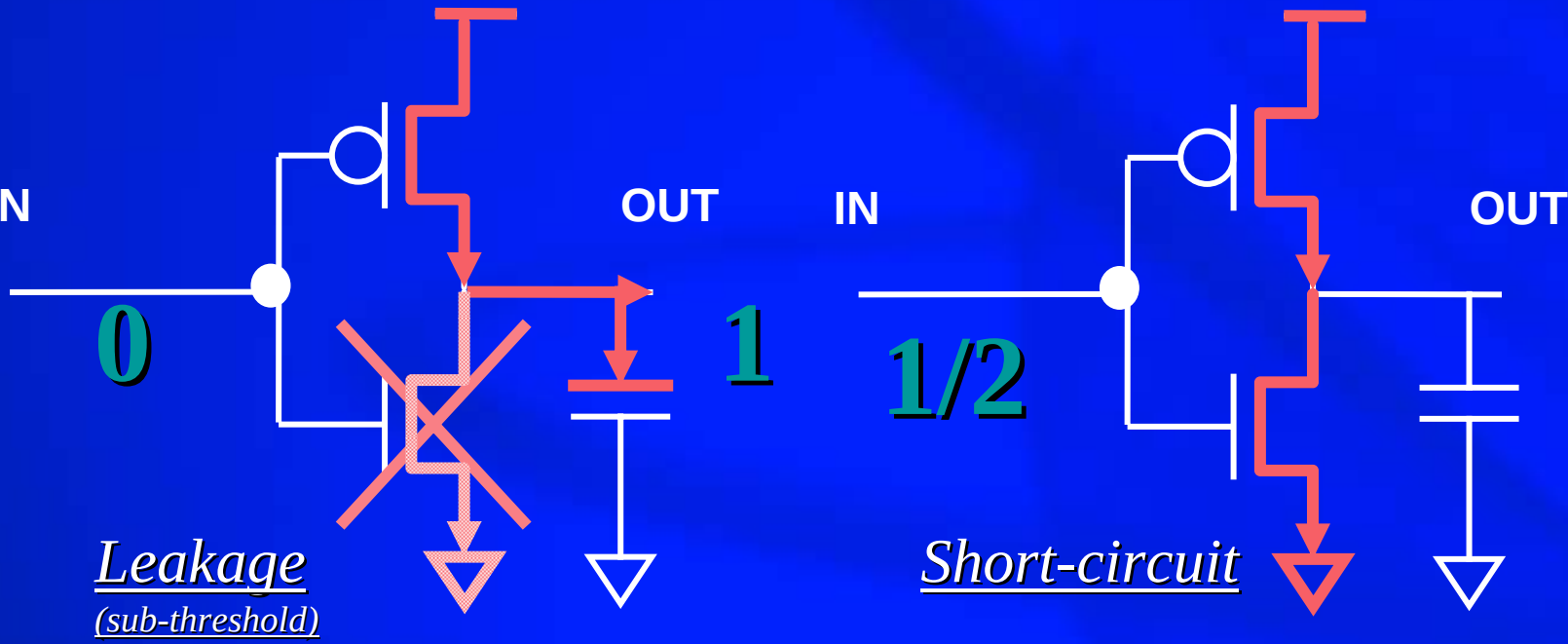




- **Power is also consumed on wires – busses, interconnects.**

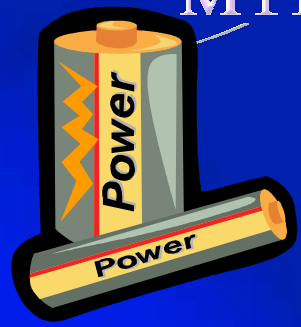
Power and the digital world (2)…

- Secondary effects like leakage and short-circuit current are increasing with advanced process technologies.



- Leakage became a 1st order factor
  - Used to be 7-15% not long ago, 20%-50% in coming generations.

Power & Energy



Energy

- Relevant in respect to:
  - Battery life - lower energy per task ➔ longer battery life.
  - Electric bills - lower energy per task ➔ lower bills.
- $E = \alpha CV^2$ - Proportional to capacitance, voltage^2, and activity

Power

- *Energy per time* - measured in Watts.
- $P = \alpha CV^2f$ (α: activity, C: capacitance, V: voltage, f: frequency)
- Peak power is a concern:
  - Higher power ➔ higher current.
  - Higher power ➔ higher temperature.

Power Density

- Think of watts/cm^2.
- Higher power, Smaller area ➔ Higher power density
- Denser power ➔ Harder to cool

Thermal

Thermal

- Power → heat → higher temperature
- Temperature increased with
  - Higher power
  - Higher power density
  - Higher ambient temperature
  - Weaker cooling (cooling = form factor, cost)
- Higher temperature risks
  - Immediate malfunctioning
  - Reduced long term reliability
  - Higher leakage power
  - Noise

Cause and Effect

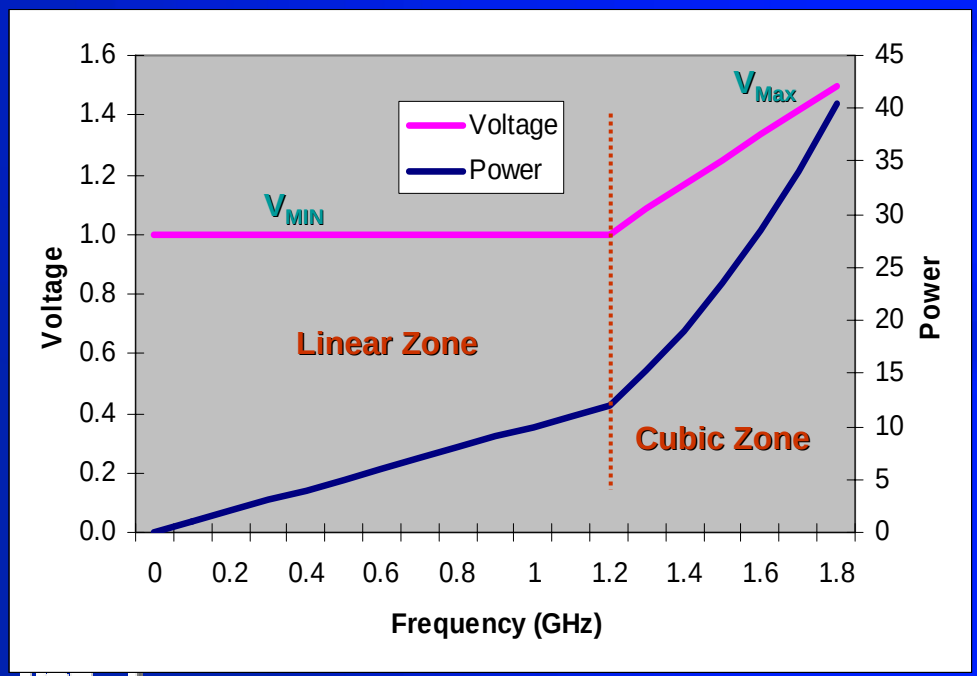
- More transistors, bigger transistors, higher voltage, higher activity
  → Higher energy, higher power
- Higher frequency
  → Higher power
- Denser logic, higher power, smaller die
  → Higher power density
- Higher power, higher power density
  → Higher temperature
- Higher temperature
  → Higher leakage power
  → Higher temperature…

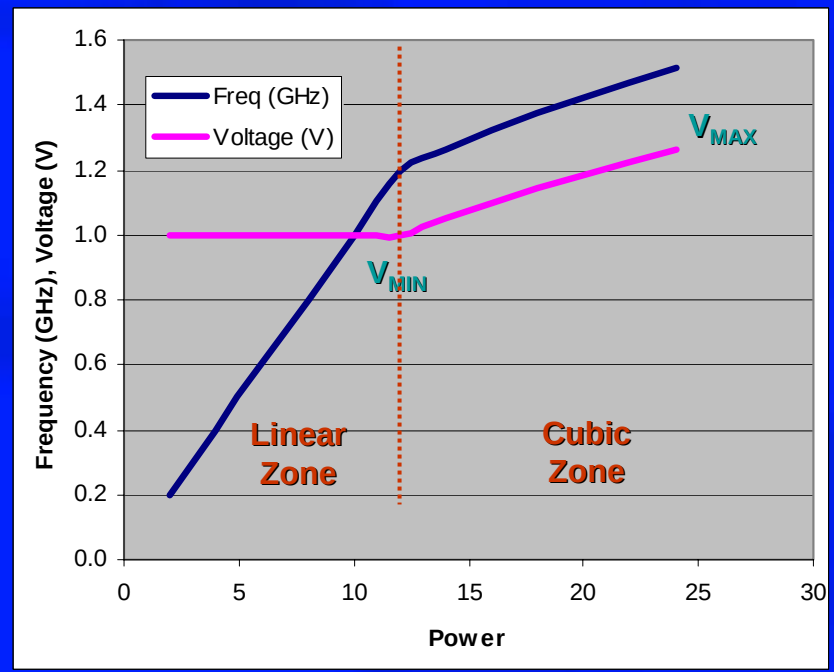
Performance, Frequency, Transistors

- Performance = IPC × Frequency
- Sources of frequency increase
  - Process: lower delay time
  - Microarchietcture: deeper pipeline
- Higher frequency → Lower IPC
  - Memory access do not scale …
  - e.g., 50% frequency gain → 35%-45% performance
- More transistors → Higher IPC. Typical#
  - Single core: 2X transistors → 1.4 IPC
  - Multi core: 2X transistor → 1.8 IPC

Performance, Power
Physics: Voltage/Frequency scaling

- *Freq ≈ k * Voltage (F ≈ kV)*
  - Within a limited voltage range V_{MIN}/V_{MAX}
- *Active Power = Activity * Capacitance * Voltage² * Freq ($P=\alpha CV^2f$)*
  - ➔ Power is proportional to f^3 ($>V_{MIN}$)
  - ➔ Power is proportional to f ($\leq V_{MIN}$)
- *Used for Static and Dynamic Voltage/Frequency) Scaling (DVS)*



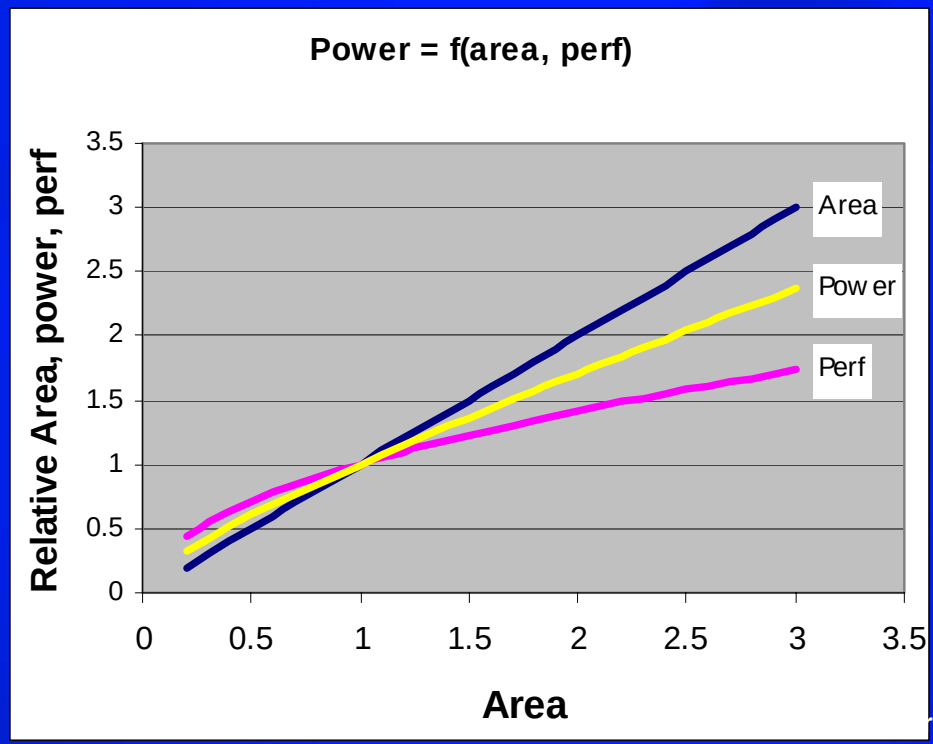


Performance, Power, Area

Empirics – the power growth

Power = f(Perf, Area)

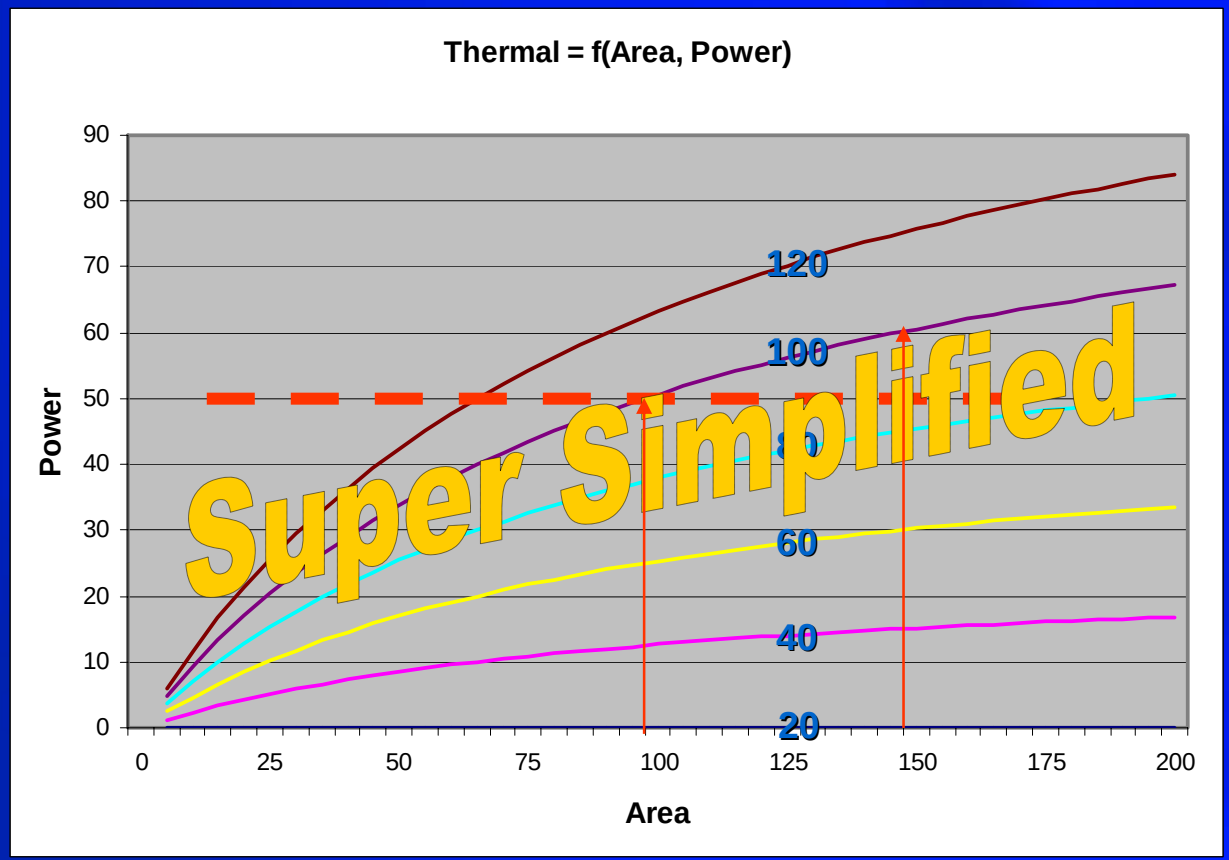
- *Capacitance = k * Area (C=kA)*
- Power Growth Guesstimate (same process, same voltage)
  - Area growth ➔ power growth
  - Performance growth ➔ power growth
  - ➔ Empirically, <u>for the same process and voltage</u>, power growth is ~ average of both (combined effect of αC)



Power, Power Density, Thermal

Thermal = f(Power, Power Density) = f(Power, Area)

- Higher Power ➔ Higher Temperature
- Smaller Area ➔ Higher Temperature



Note:
Same power
smaller area
➔ Higher temp.

Qualitative Data
Assumes ambient temp. 20°

Process Technology – In Practice

Every new process technology (~2-3 years cycle) changes the physical attributes of the transistor:

- Size reduced to 0.5X
- Delay time reduced to 0.75X
- Switch energy reduced by 1.75X (Relative to CV^2)
  - Capacitance per transistor reduced to 0.7X
  - Operation voltage reduced to 0.9X

Two extreme scenarios:

- Ideal "Shrink"
  - ***Same μarch***

1X	#Xistors
0.5X	size (0.7X per dimension
1.35X	frequency
0.9X[1]	IPC
0.75X	peak power
1.2X	performance (IPC×freq)
1.5X	power density

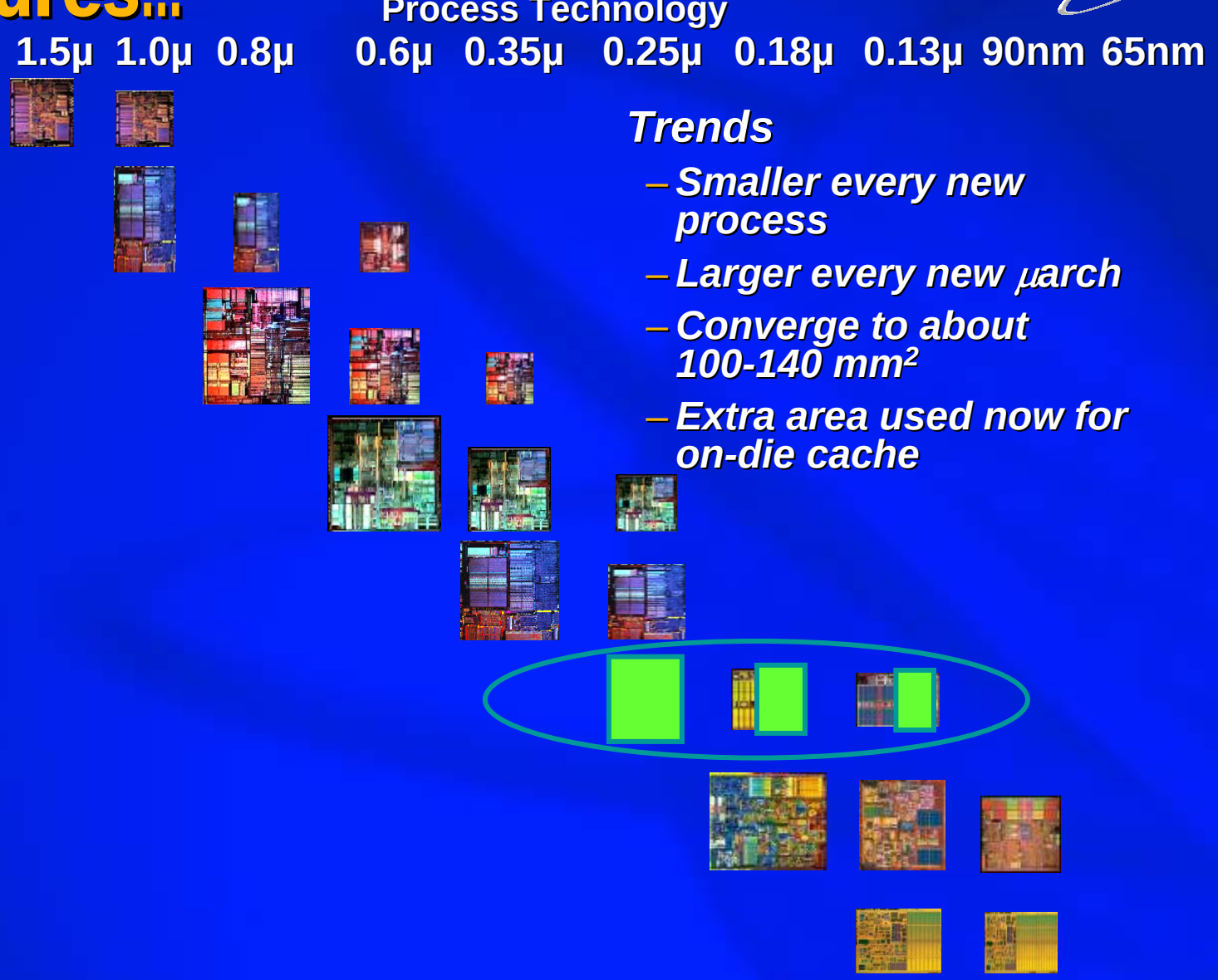
- Ideal New design
  - ***Same die size***

2X	#Xistors
1X	size
1.35X	frequency
1.3X-1.7X[1]	IPC
1.5X	peak power
1.8-2.3X	performance
1.5X	power density

[1] IPC Decreased in higher frequencies

Still Good – but less than we hoped for

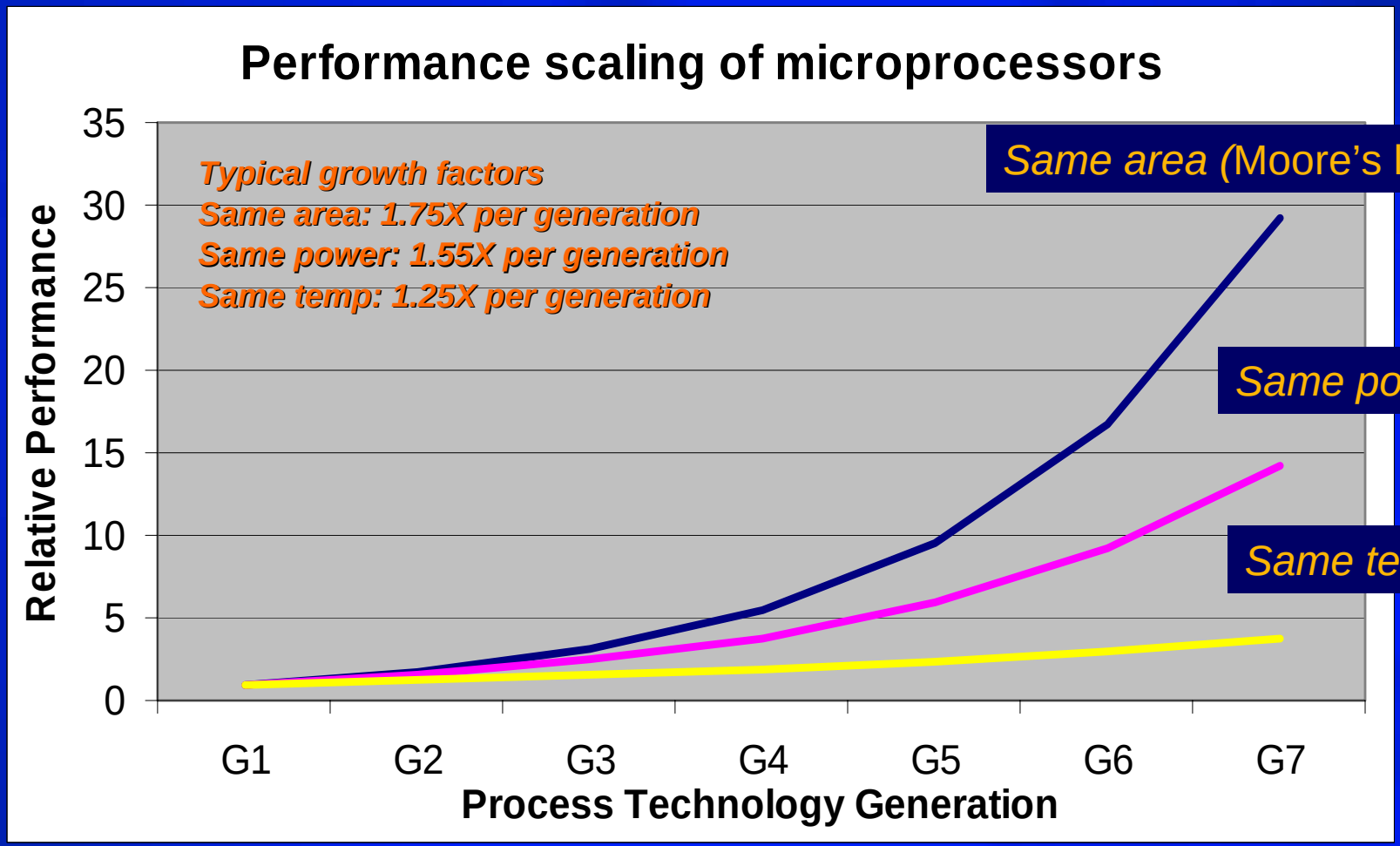
In Pictures...



Trends

- *Smaller every new process*
- *Larger every new µarch*
- *Converge to about 100-140 mm²*
- *Extra area used now for on-die cache*

The Thermal Wall

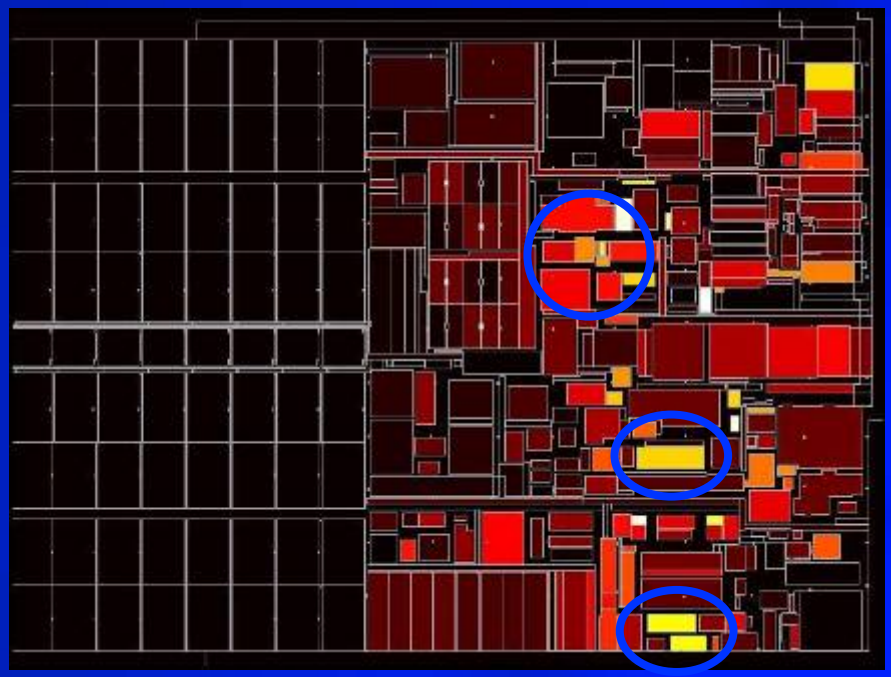


In a thermally limited environment
Evolutionary Uarch will diminish its performance return

And it Gets Even Worse…

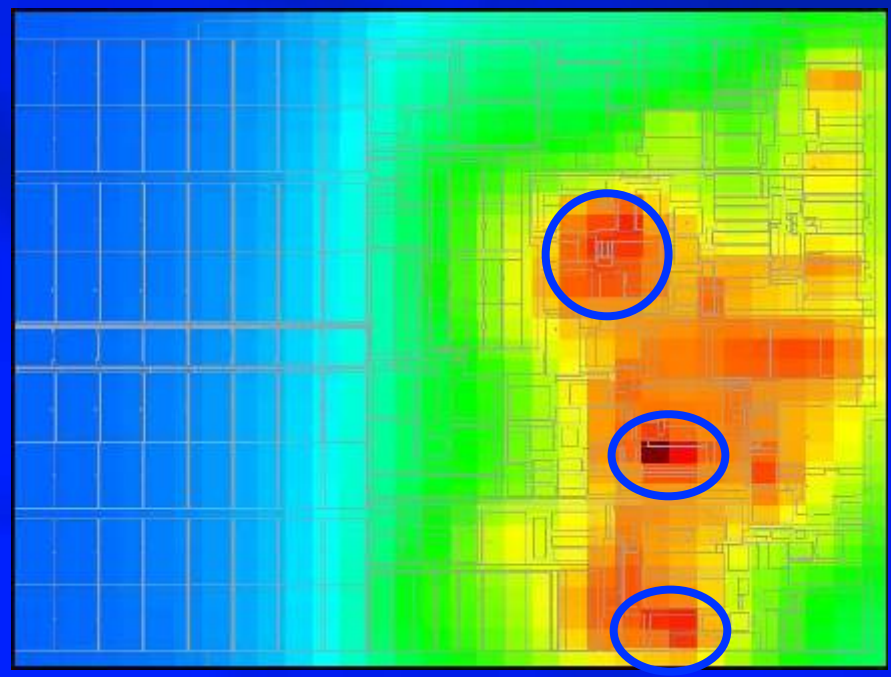
- Variation in space
- Variation in time
- Variation in process
- Variation in platform (form factor)
- Variation in cooling solutions
- OS dependence (ACPI)

Thermal Variation - Space



Power Density (simulated)[1]

Color codes: (lowest) black, red, orange yellow, white (highest)



Thermal Map (simulated)[2]

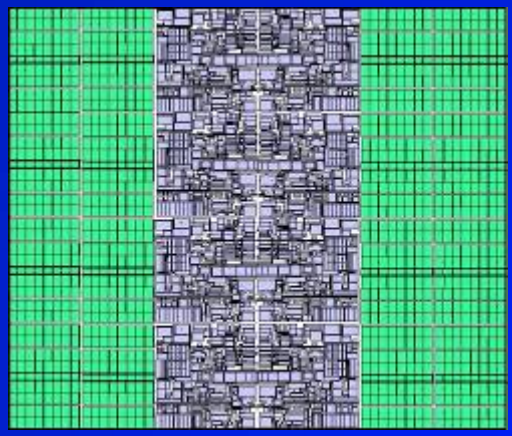
(lowest) blue, green, yellow, orange, red (highest)

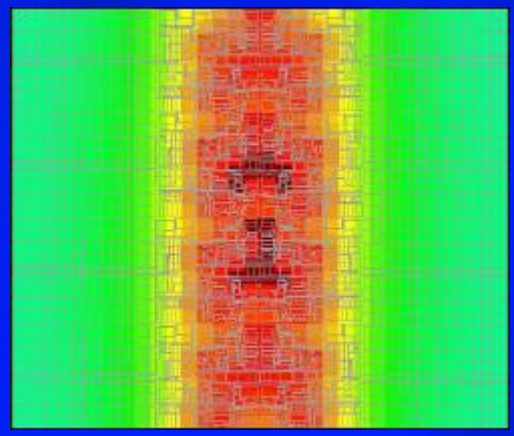
Pentium M Processor power density example

[1] Source : Intel® Pentium® M Processor Power Estimation, Budgeting, Optimization, and Validation
Dani Genossar, Nachum Shamir, ITJ Q2/2003

[2] Source: Lev Finkelstein, Intel 2005

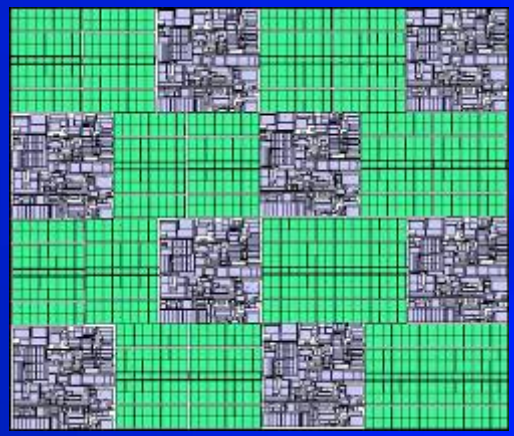
Thermal Variation – Space / CMP

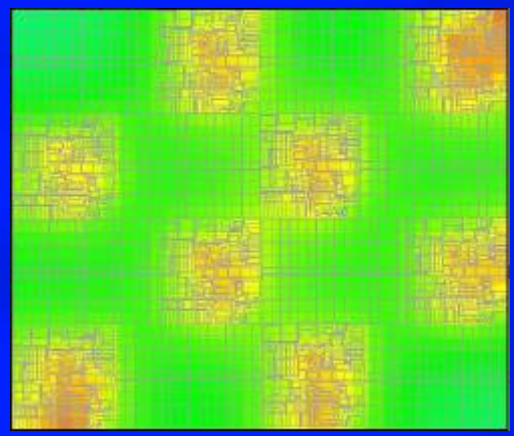




8 cores on the center
Tjmax: 100.4°C

Qualitative Data



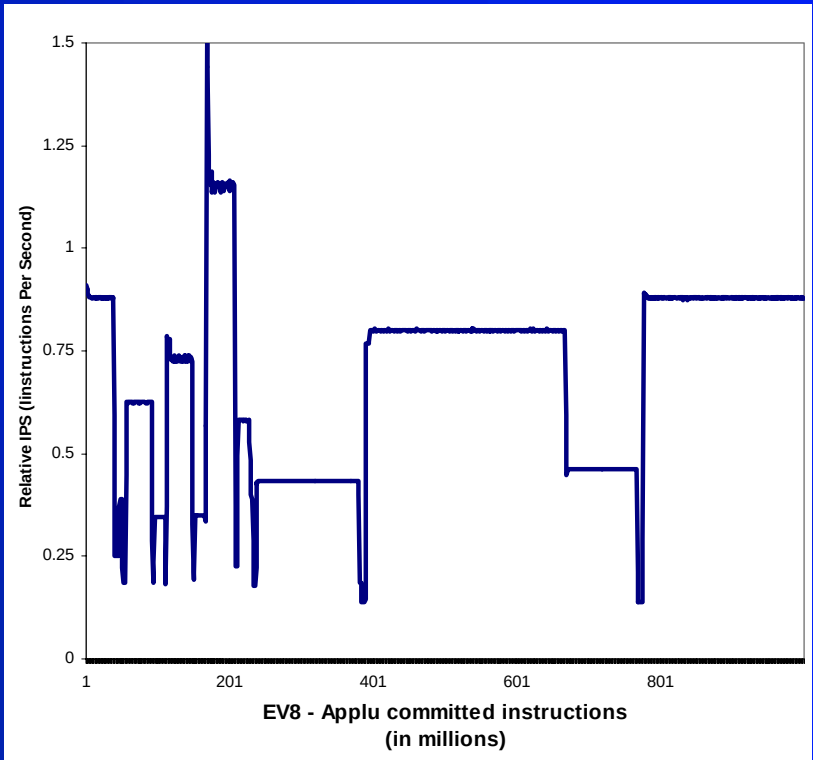


8 cores checkers layout
Tjmax: 96.3°C

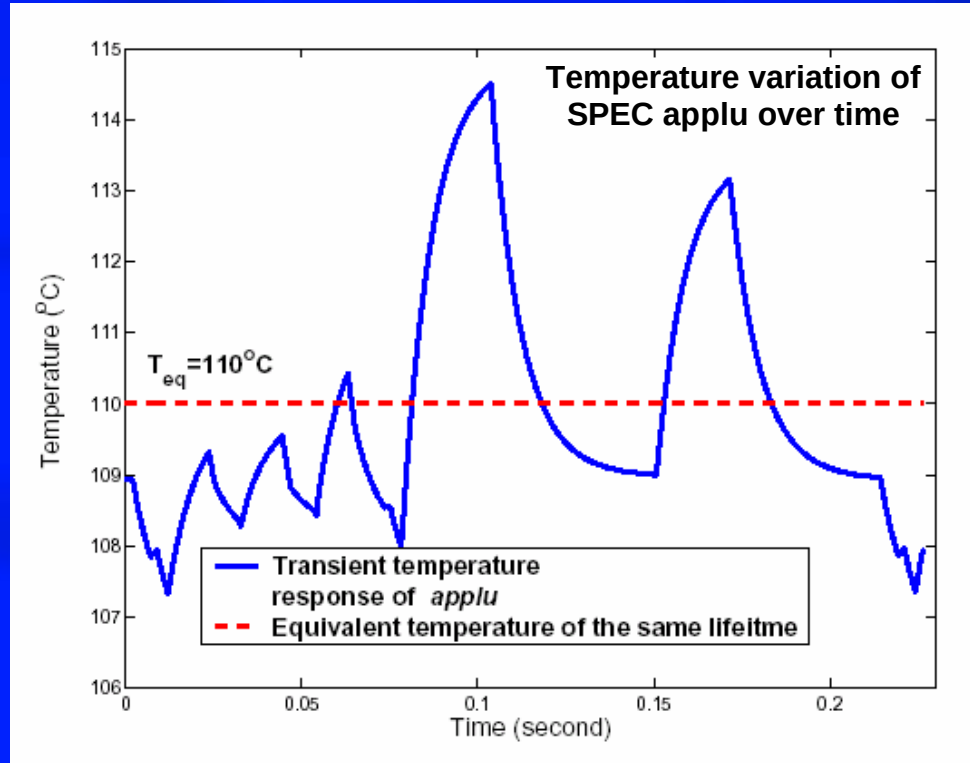
- Uniform power distribution is better
- Chip Multi-Processor (CMP) can exploit that

Thermal Variation – Time

- Follows activity ➔ tracks IPC
- Varies among applications
- Varies within an application



Source: "Processor Power Reduction Via Single-ISA Heterogeneous Multi-Core Architectures" Rakesh Kumar et.al Micro'03



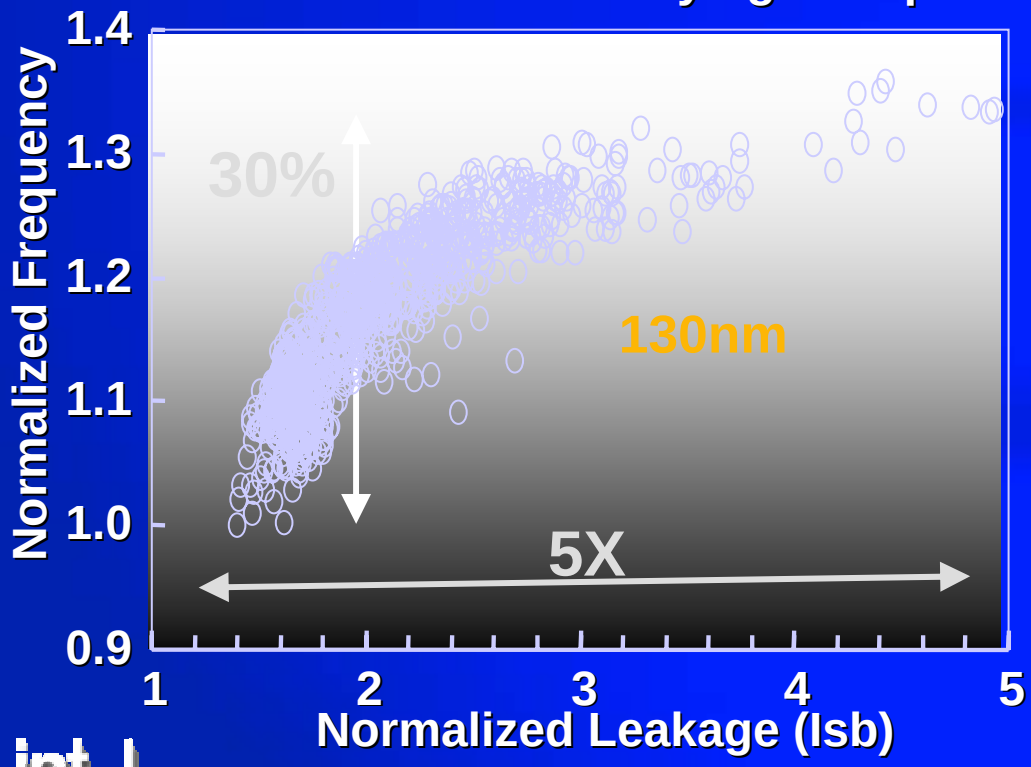
Source: "A Quick Thermal Tutorial" Kevin Skadron, Mircea Stan, U. of Virginia 2005

Two graphs are not from exact same code

Thermal Variation - Process

- Random dopant fluctuations
- Sub-wavelength Lithography

➔ Varying feature size
➔ Varying delay time
➔ Varying leakage power
➔ Varying total power, temprature

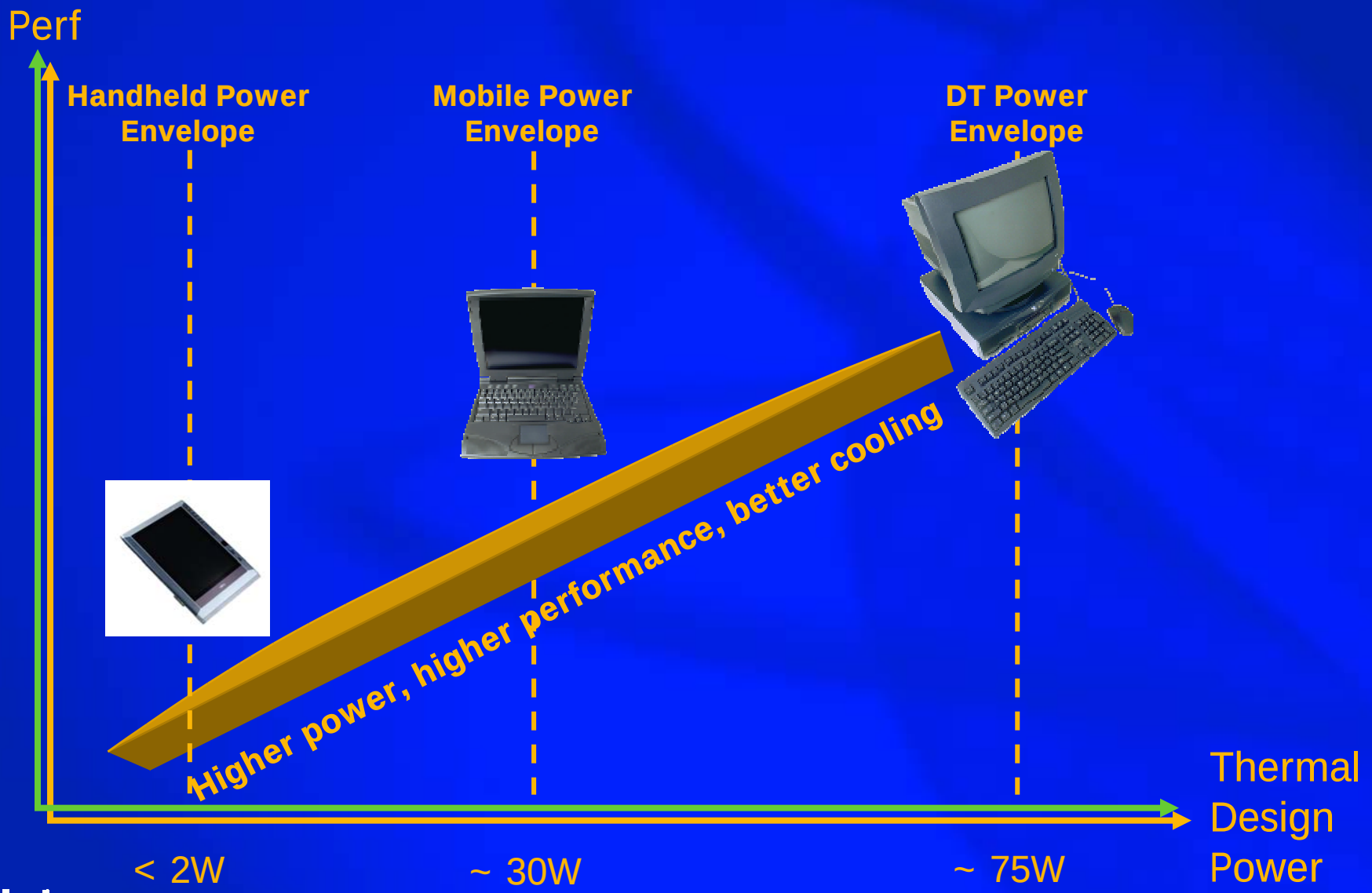


Frequency
~30%

Leakage Power
~5-10X

Source: Microarchitecture and Design Challenges for Gigascale Integration
Shekhar Borkar, Intel,
Micro-37 Keynote 12/2004

Thermal Variation - Platform



Implications

- Thermal events no longer a rare exception
  - Can happen frequently
  - In different scenarios

- Worst-case design is extremely inefficient and error prone

Approach:

- *Dynamic Thermal Management*

Dynamic Thermal Management (DTM)

- Basics
- “Linear” Thermal Throttling
- DVS* based Thermal Throttling
- Micro-architectural Thermal Throttling
- Advanced DTM
- Optimal DTM

*DVS: Dynamic Voltage Scaling

Thermal Management - Basics

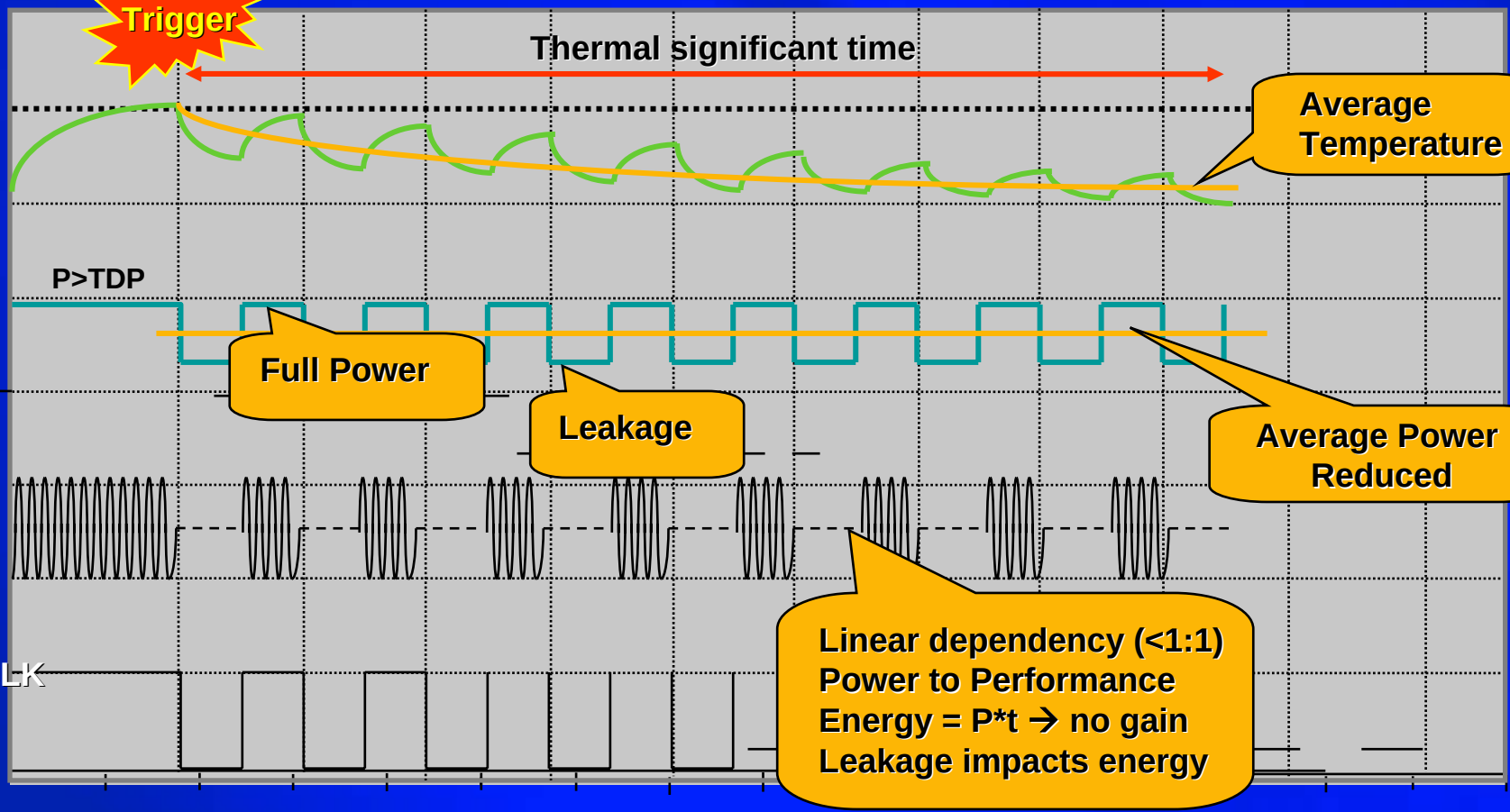
- Avoid over heating by reduction in activity
- Triggered by thermal event
  - Actual or predicted
  - Traditionally detected by thermal sensor

- Activity reduction methods
  - “Stop clock”
  - Frequency reduction
  - Voltage/frequency scaling, …

- When to throttle
- Where to throttle
- For how long? (when to stop)

“Linear” Thermal Throttling

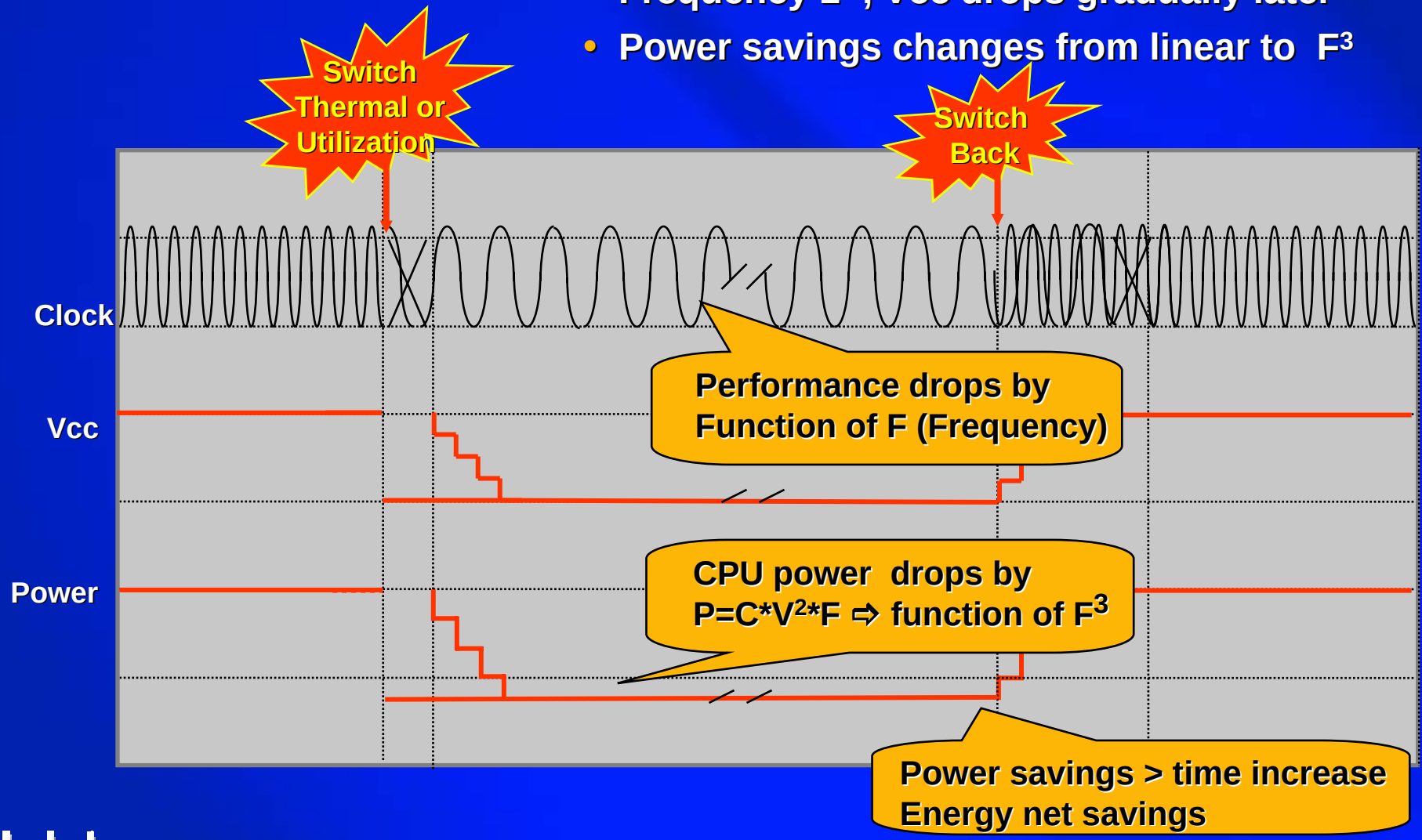
- Stop all activity for a predefined time
- Power & Temperature down, energy up



Source: Analysis of The Enhanced Intel® Speedstep® Technology of the Pentium® M Processor, Efi Rotem et al TACS-1 6/2004

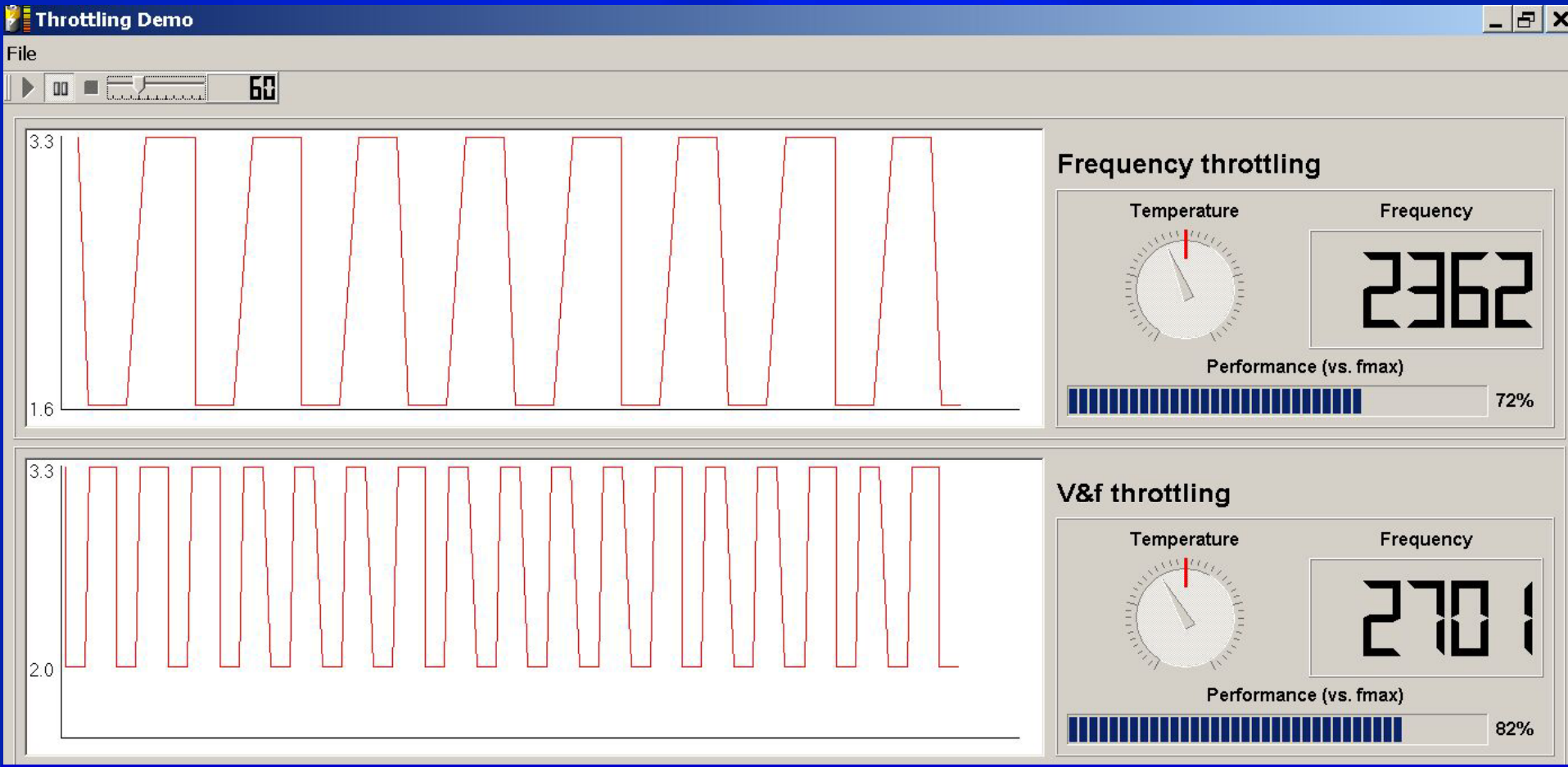
DVS based Thermal Throttling

- Reduce both Voltage & Frequency
- Frequency 1st, Vcc drops gradually later
- Power savings changes from linear to F^3



Source: Analysis of The Enhanced Intel® Speedstep® Technology of the Pentium® M Processor, Efi Rotem et al TACS-1 6/2004

DVS vs. Linear Throttling



DVS Thermal Throttling – The Challenge

- Maximize performance for given thermal environment

Constraints

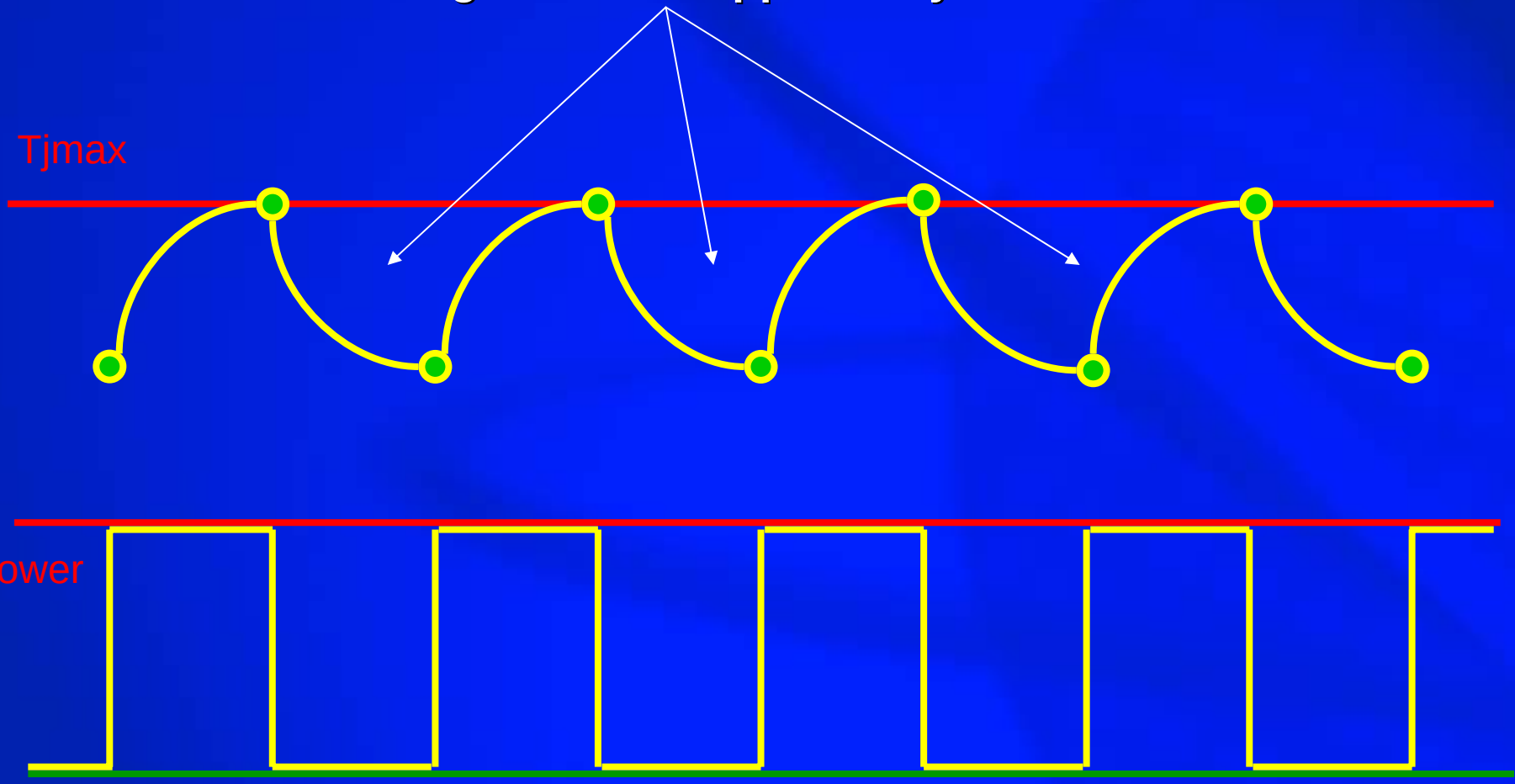
- Voltage/frequency changes take time
- Longer operation at lower frequency ➔ lower performance

Questions

- When to throttle?
  - Should we wait for the last moment?
- Where to throttle?
  - Should we go for lowest voltage/frequency?
- When to stop throttling?
  - How much time should we cool down? *Sprint or Marathon?*
  - To which voltage/frequency level?

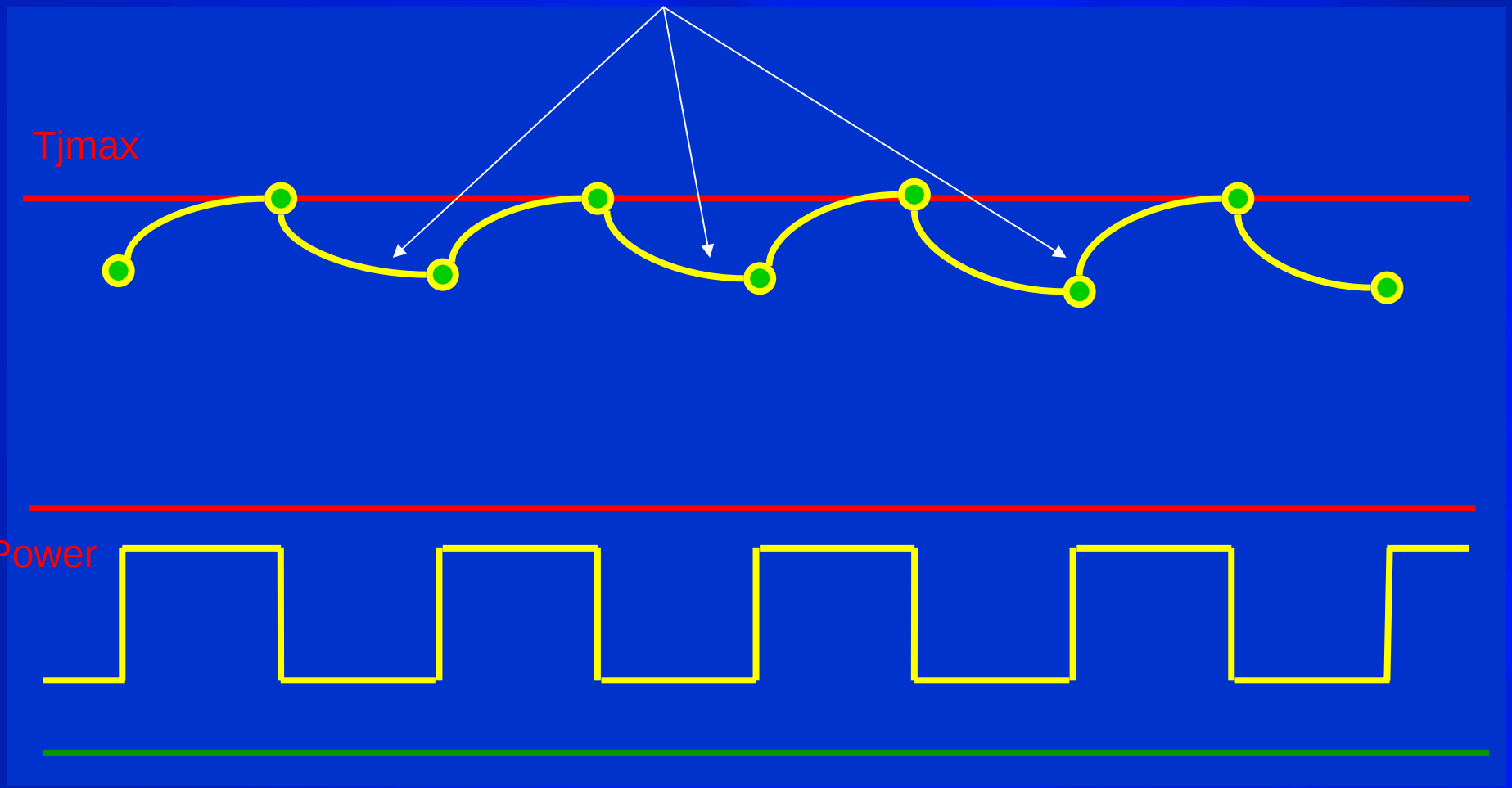
Where to Throttle?

Region of lost opportunity



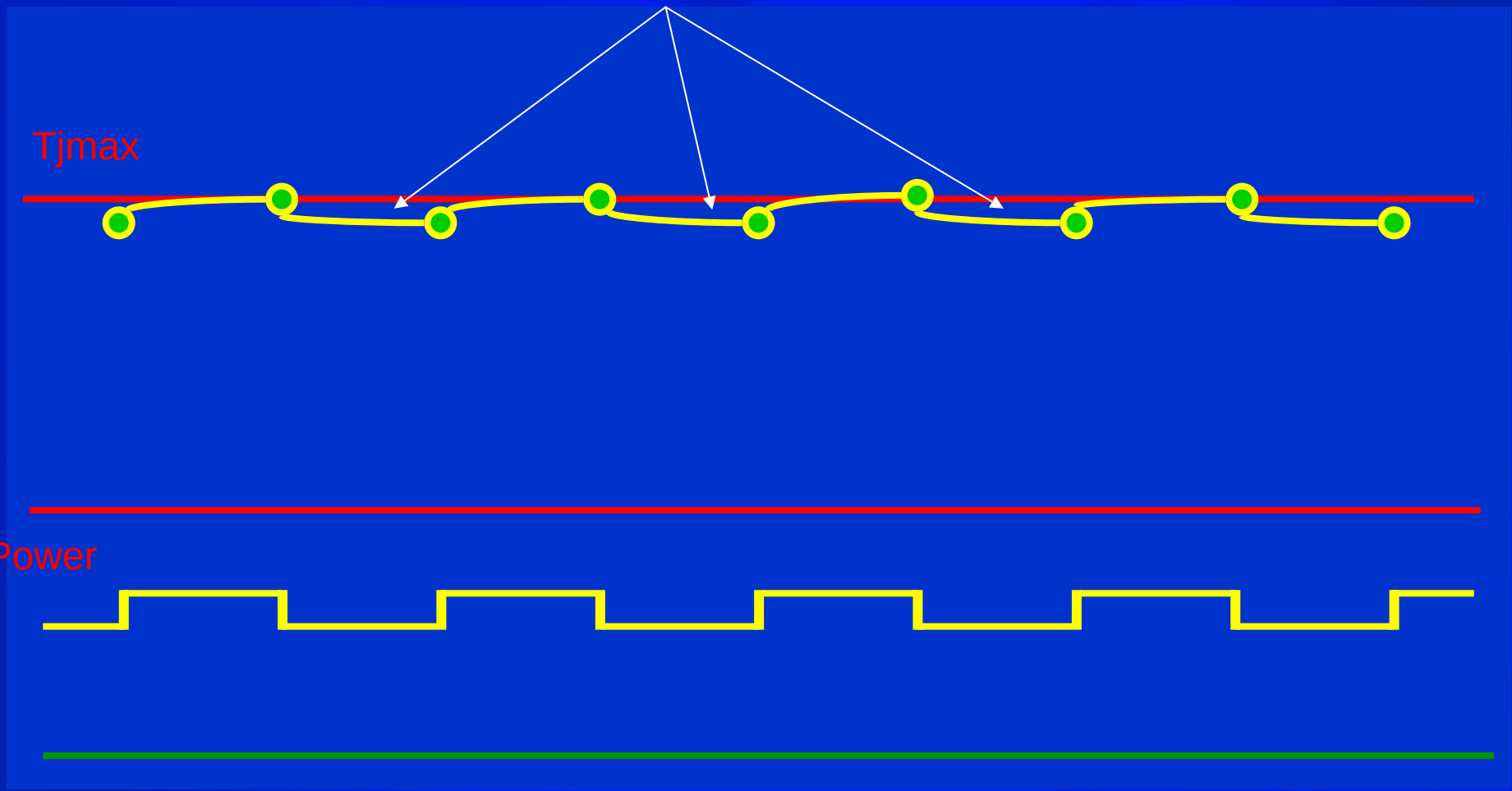
Where to Throttle?

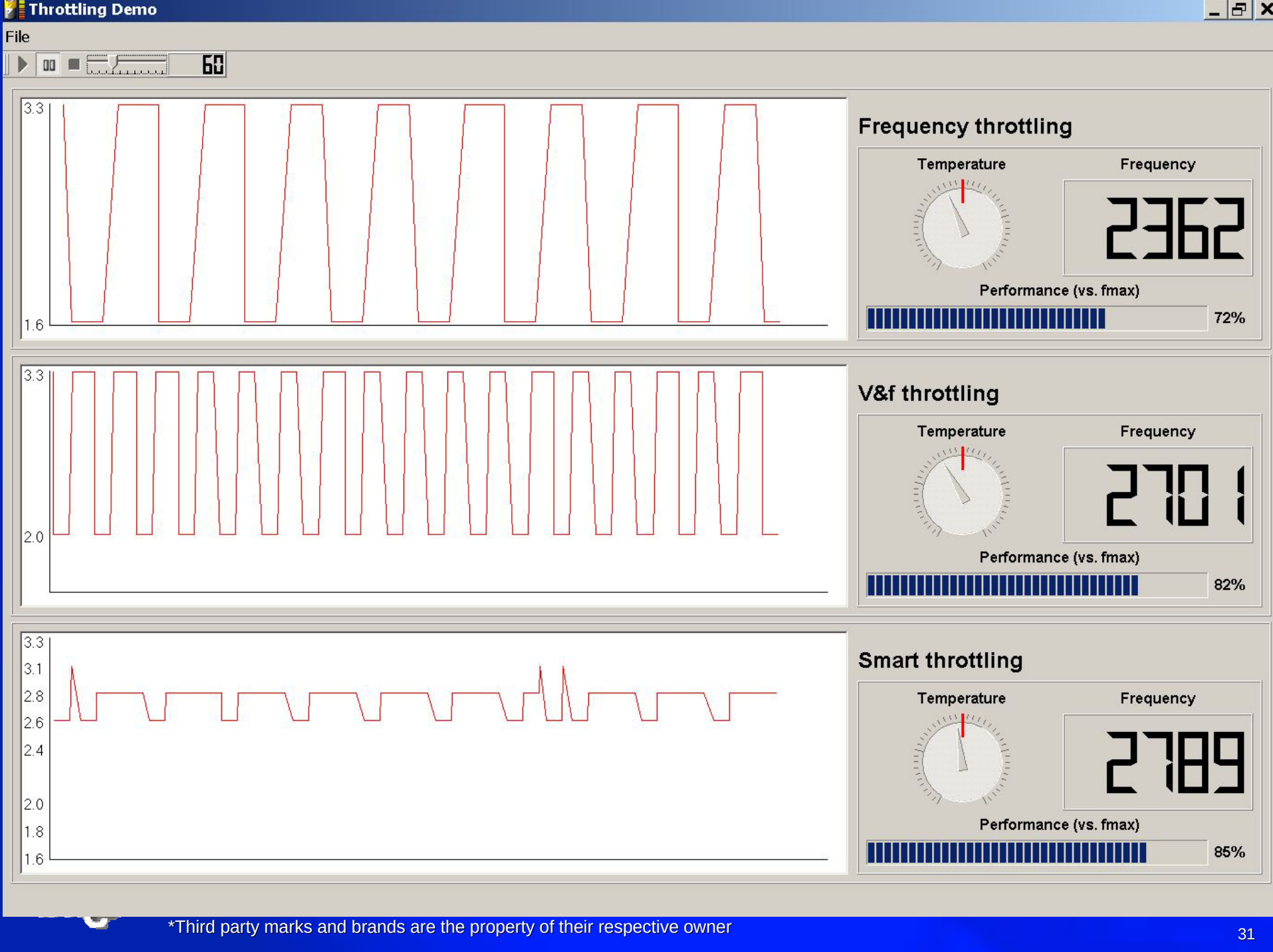
Somewhat Better



Where to Throttle?

Much Better





Micro-Architectural Throttling

- Reduce activity by slowing certain units
- Reduces power → reduces temperature
- Examples:
  - Stop or slow down instruction fetch[1]
  - Stop speculation
  - Adaptively reduce array sizes
- Advantages
  - Fast activation
  - Super-linear savings
  - Focused targets
- Disadvantages
  - Usually, less efficient than DVS

[1] Pipeline Gating:
Speculation Control for Energy Reduction. Srilatha Manne, Artur Klauser, Dirk Grunwald, ISCA '98

Advanced DTM

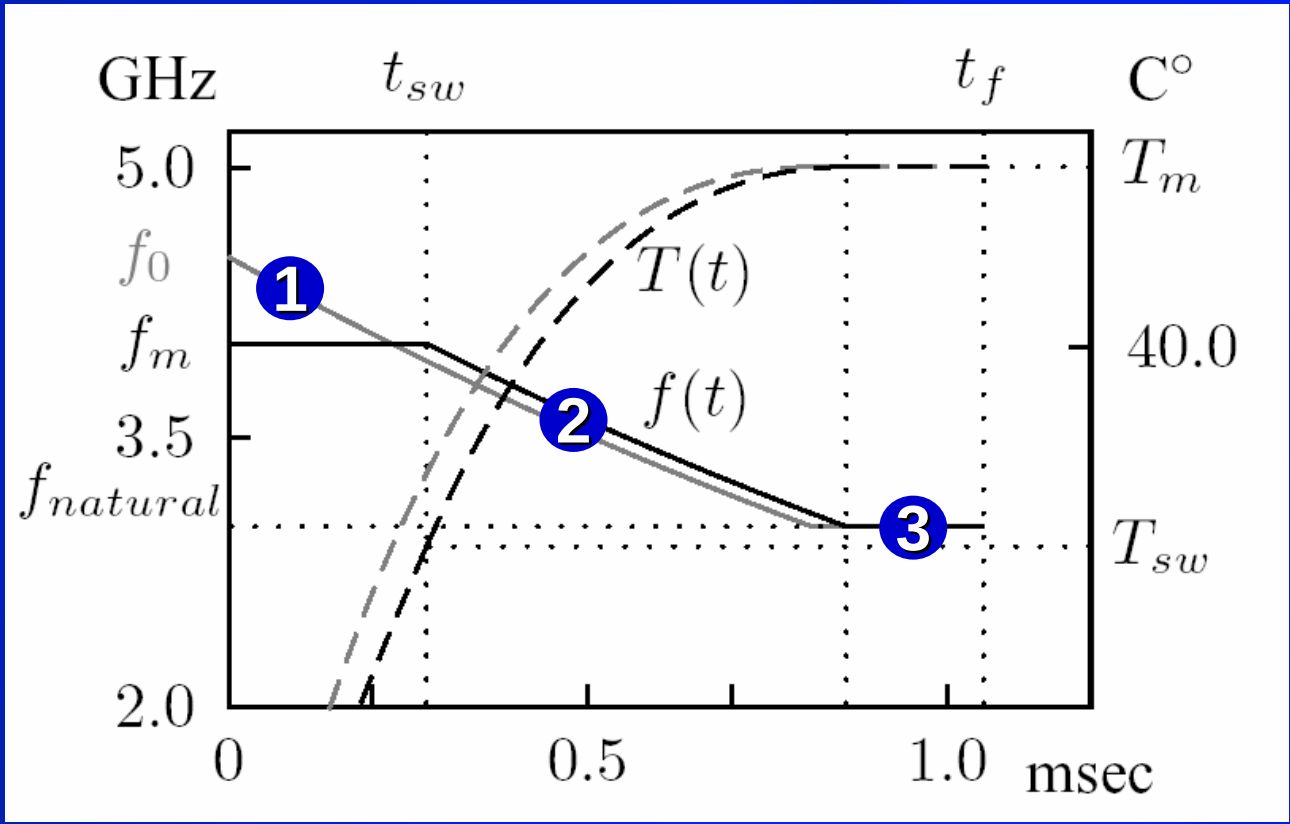
- **Goal:**
  Maximize performance per thermal constraints

- **Reactive systems: fast, minimal input,**
  - **Throttling: on/off scheme, thermal interrupts.**
- **Controllers**
  - **PID[1] control - tune power to keep temperature**
    - **Fast, minimal history usage, needs power estimation**
- **Even smarter mechanisms**
  - **Predictive techniques**
    use history to predict power & temperature behavior
- **Tradeoff**
  input type/size ⇔ computational efforts ⇔ reaction time ⇔ quality

[1] **Proportional, integral, Differential**

"Optimal" Control – example

Maximal Performance per Given Max Tj



1. Can run higher than maximal "natural" frequency and can dissipate more than maximal power for limited time
2. Start decreasing power even before reaching Max Tj
3. Eventually approach "natural" maximal power & frequency

Optimal strategy for a single-RC model[1] (Starting at an ambient temperature)

[1]From Cohen et. al, "On Estimating Optimal Performance of CPU Dynamic Thermal Management." Computer Architecture Letters, Volume 2, Oct. 2003

Summary: Observations

- Processors are becoming thermally limited
  - It is 1st order problem … and will just get worse.
- Thermal solutions improve –
  But cost and form factor limit their benefit
- Variations prohibit worst case design
  - Too inefficient
  - Error-prone
- Necessitates dynamic thermal management
- DTM evolution – from simple-reactive to smart-proactive

Summary: Challenges

- Reduce Power/Temperature
  - To actively eliminate/relax heating
- Improve measurements
  - To know better where we are
- Improve control
  - To get the maximum in a given environment
- Architecture breakthrough
  - Significant improvement in power-awareness

The End

Backup Slides

General trends

Process Technology:

- ☺ Smaller (2x), faster (1.35x) transistors, consuming less energy per switch (1.75x) – but:
- ☹ More transistors per mm^2, More switches per cycle ➔ More power density (~1.5X)

Micro-architecture:

- ☹ Traditional - More transistor switches per inst.
- 😐 New world - More efficiency ➔ More activity per area
- ➔ Higher power, higher power density
- ➔ Higher temperature

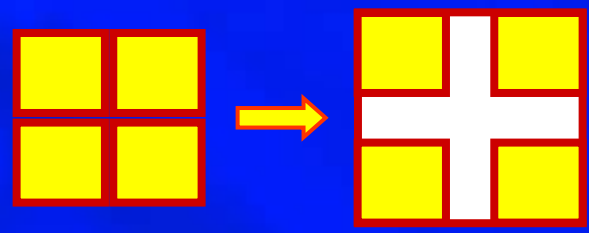
The Thermal Wall

With Naïve Process Technology:

- ☺ **Same Area**
  ➔ 2X transistors, 1.35X frequency
  ➔ 1.8X-2.3X perf., 1.3X-1.5X Power, 1X Area
- 😐 **Same Power ➔**
  1.33X transistors, 1.35 frequency
  ➔ 1.5X-1.8X perf., 1X Power, 0.67X Area
- ☺ **Same Thermal ➔**
  1.33X transistors, 0.8X power
  ➔ 1.25X frequency, 1.35X-1.65X perf…

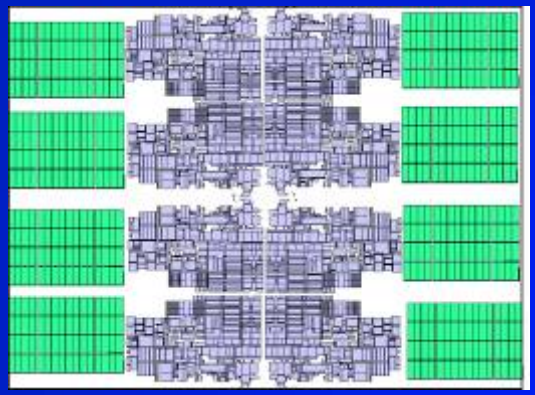
Simple Solutions

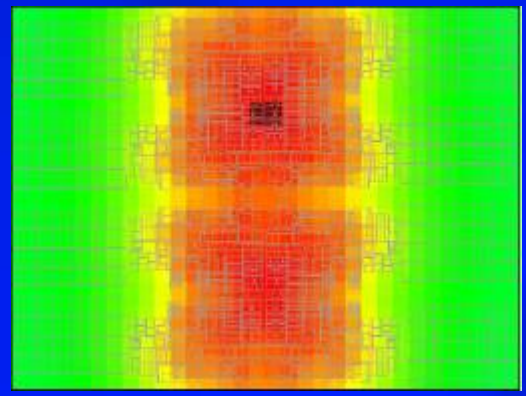
- Spread the transistors over the area
  - ☺ Reduce power density
  - ☹ Higher wire delays



- Use voltage scaling
  - ☺ Reduce power & power density
  - ☺ Allow more transistors
  - ☹ Lower frequency → Lower performance

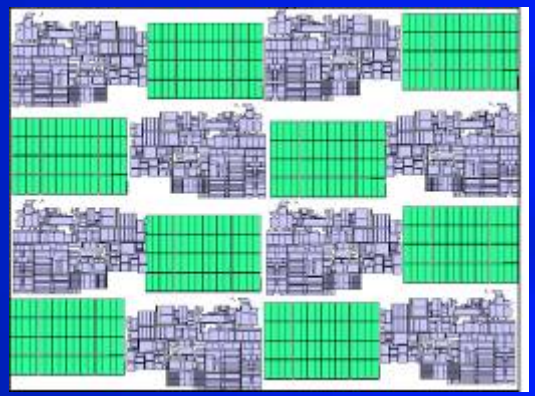
Thermal Variation – Space / CMP

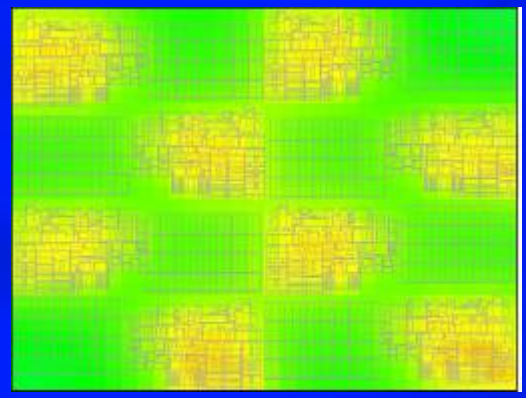




8 cores on the center
Tjmax: 100.4°C, leakage: 19.3W

Qualitative Data





8 cores checkers layout
Tjmax: 94.0°C, leakage: 18.06W

- Uniform power distribution is better
- Chip Multi-Processor (CMP) can exploit that

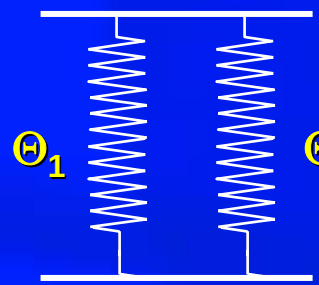
r22 Replace w/ Banias

ronen; 27/4/2005

Thermal Basics

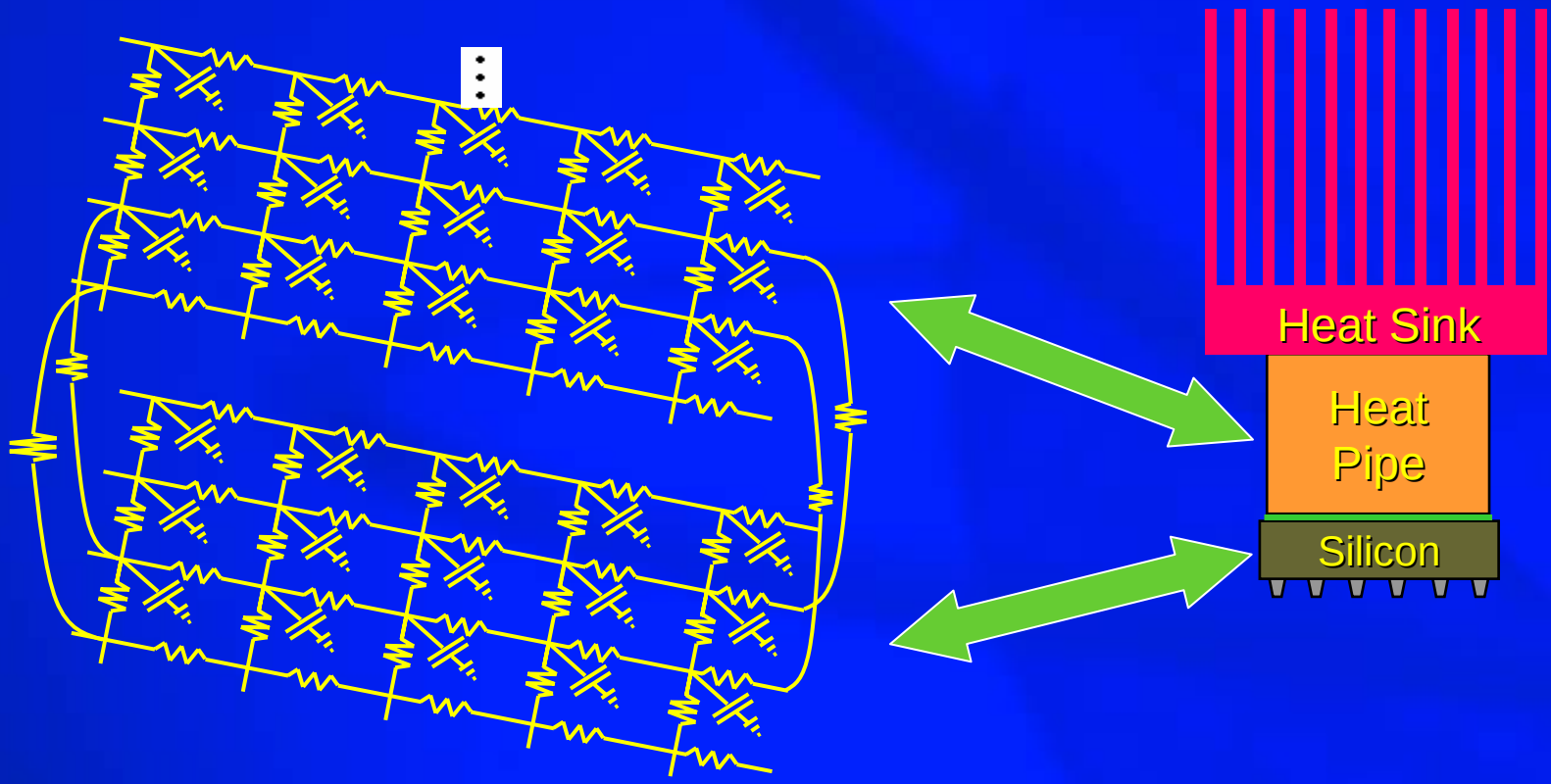
- $\Delta T = P * \Theta$
  - Θ is "Thermal resistance"
  - Similar to $\Delta V = I * R$
  - T is measured in ^oC, P in Watts, Θ in $^oC/W$.
- As in resistors we see:
  - Serial: $\Theta = \Theta_1 + \Theta_2$
  - Parallel: $1/\Theta = 1/\Theta_1 + 1/\Theta_2$
- Θ behaves like resistance
  - Grows with length
  - Reduced with area
  - ➔ For a piece of silicon, we can measure R = Θ per area: $^oC/(W*cm^2)$
- ➔ Simplified model
  - ➔ Realistic models take into account power distribution and thermal maps

$P = P_1 = P_2$
$\Theta = \Theta_1 + \Theta_2$
$\Delta T = \Delta T_1 + \Delta T_2$



$P = P_1 + P_2$
$1/\Theta = 1/\Theta_1 + 1/\Theta_2$
$\Delta T = \Delta T_1 = \Delta T_2$

RC networks for Thermal



Challenges – Improve Measurements

Simulation

- Better, more accurate, thermal estimate
  - More detailed and accurate power modeling
  - Integrated active/leakage with thermal modeling
  - Accurate thermal solutions

On chip

- Temperature measurement on die.
  - More accurate, more points on chip
  - Targeting hot spots

Challenges – Reduce Temperature

Power Awareness

- ***"Less is More"***
  Do the same "work" with less power
  - **Less instruction per task**
  - **Less micro-operation per instruction**
  - **Less transistor switches per micro-operation**
  - **Less energy per transistor switch**
- **Better power/performance trade off**
  - **Optimal/adaptive structure sizes, …**

Reduce power density

- **Identify hot spots – distribute them as possible**

Better cooling solutions

Challenges – Improve Control

Adaptive Thermal management

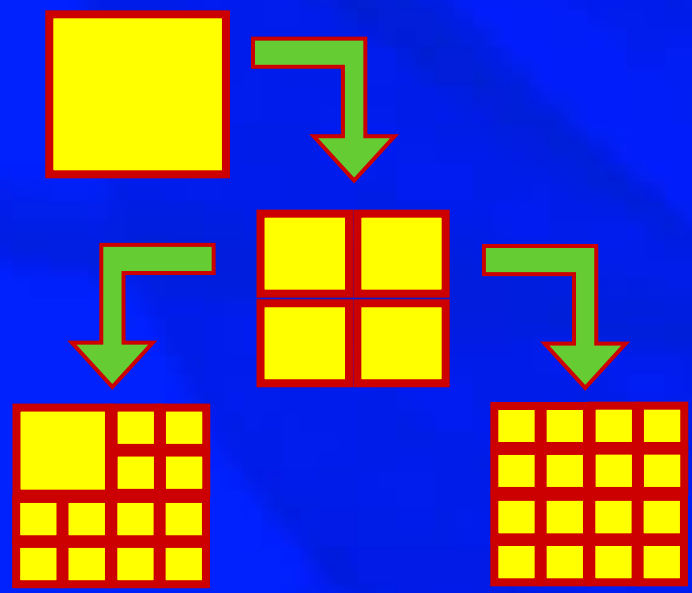
- Eliminate risk of thermal run-away
  - Already being done in today's processor
- Maximize performance in thermally limited environment
  - Technology in infancy
  - Lack of determinism is a concern for OEMs*
- This is an optimization problem!
  - How to control power ➔ thermal for best performance
  - Current Challenges: SMT/MP.

* See http://www.vanshardware.com/articles/2004/05/040517_efficeonFreeze/040517_efficeonFreeze.htm

What next? Asymmetric Cores?

- Mix big and small cores
  - Big core(s) – for single thread performance
  - Small cores – for <u>efficient</u> multithreaded performance*

	ST perf	MT perf
1 core	1X	1X
4 cores	½X	2X
16 cores	¼X	4X
13 cores	½X	3.5X



Best of all worlds?

* Assume smaller core has ¼ area, ¼ power and ½ performance